Stochastic comparisons of stratified sampling techniques for some Monte Carlo estimators

Larry Goldstein
Department of Mathematics
University of Southern California
Kaprielian Hall, Room 108
3620 Vermont Avenue
Los Angeles, CA 90089-2532, USA
larry@math.usc.edu

Yosef Rinott*
Department of Statistics
and Center for the Study of Rationality
Hebrew University of Jerusalem
Mount Scopus
Jerusalem 91905, Israel
and LUISS, Roma
rinott@mscc.huji.ac.il

Marco Scarsini†
Dipartimento di Scienze Economiche e Aziendali
LUISS
Viale Romania 12
I–00197 Roma, Italy
and HEC, Paris
marco.scarsini@luiss.it

April 2009

*Partially supported by the Israel Science Foundation grant No. 473/04
†Partially supported by MIUR-COFIN

Abstract

We compare estimators of the (essential) supremum and the integral of a function f defined on a measurable space when f may be observed at a sample of points in its domain, possibly with error. The estimators compared vary in their levels of stratification of the domain, with the result that more refined stratification is better with respect to different criteria. The emphasis is on criteria related to stochastic orders. For example, rather than compare estimators of the integral of f by their variances (for unbiased estimators), or mean square error, we attempt the stronger comparison of convex order when possible. For the supremum the criterion is based on the stochastic order of estimators.

For some of the results no regularity assumptions for f are needed, while for others we assume that f is monotone on an appropriate domain. Along the way we prove convex order inequalities that are of interest *per se*.

AMS 2000 Subject Classification: Primary 65C05. Secondary 60E15, 62D05.

Keywords: convex order, stochastic order, majorization, stratified sampling.

1 Introduction

In many situations the cost of computing the value of a function at some point is very high, either because the analytic expression of the function is extremely complex, or because the value is the result of an experiment. Therefore, due to budget restrictions, the function can be computed only at a finite number of points. Often the object of interest is not the whole graph of the function, but only some functional. Monte Carlo estimation of functionals such as the maximum or the integral of a real valued function f is the subject of a very large number of papers. In most cases some regularity of the function f is assumed, see, for example, Novak (1988) or Zhigljavsky and Chekmasov (1996). Moreover, in much of the literature, estimators are compared in terms of a given loss function, which may be arbitrary. The emphasis in this paper is on showing that more general comparisons of estimation methods are possible in terms of suitable stochastic orders that imply comparisons for wide classes of loss functions. At the same time, we attempt to work with minimal assumptions on f. In this paper we will compare estimators with respect to different stratified sampling schemes, and will show that, generally speaking, refining stratification leads to an improvement of estimators.

We briefly remark that the results can be interpreted also in terms of finite population sampling. One advantage of the framework used in this paper is that finite population correction factors do not appear.

Our interest in the topic started with the result given below in Corollary 3.3, that we heard from Abram Kagan in 2006, without reference; the only written statement and proof that we were able to find (much later) is in Zhigljavsky and Žilinskas (2008). For a measureable function f defined on an interval J consider the following two estimators X and Y of the supremum of f. To compute X, sample n points $U_1, \dots, U_n$ uniformly from J and set $X = \max(f(U_1), \dots, f(U_n))$. For Y split J into n subintervals of equal length and let $V_1, \dots, V_n$ be observations taken uniformly from each subinterval. Set $Y = \max(f(V_1), \dots, f(V_n))$. The result states that Y stochastically dominates X. Since both estimators underestimate the supremum of f, Y is clearly preferable to X as an estimator of this supremum. No regularity assumption is required for f.

Zhigljavsky and Žilinskas (2008) provide the following generalization of the above result for a function f defined on any measurable space: if we consider two partitions of the space, one of which is a refinement of the other, and we sample in proportion to the measure of each element of the partition, the more refined partition produces a stochastically larger estimator of the supremum. In our paper we generalize their results and show that the stochastic comparison for estimators of the supremum holds also when observations are censored, that is, when for a sample of pairs of random variables (U_i, Z_i) we only know whether $Z_i \leq f(U_i)$ or not.

The other focus of this paper is the estimation of the integral of a function f with respect to some probability measure, or in other words, the estimation of the expectation $\mathbb{E}[f(U)]$ for some random variable U, under no additional assumptions on f other than the existence of $\mathbb{E}[f(U)]$. We consider unbiased estimators based on sampling according to partitions as for the supremum, and compare them in terms of their variances or, when possible, in terms of the convex order. The stronger convex order comparison is possible either when the function f is monotonic, or when observations are censored. We also consider situations where the function f can only be observed at the sampled points with noise.

The paper is organized as follows. Section 2 fixes notation and reviews various properties of stochastic orders and certain dependence structures. Section 3 compares estimators of the supremum of a function based on a sample of its values. Section 4 deals with the same problem when observations are censored. Section 5 compares estimators of integrals. Section 6 does the same when observations are censored. Section 7 considers monotone functions. Finally Section 8 studies symmetric sampling designs for the estimation of $\mathbb{E}[f(U)]$.

2 Notation and preliminaries

In the whole paper a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ is assumed in the background. The *stochastic order* $\le_{\mathrm{st}}$, the *convex order* $\le_{\mathrm{cx}}$, the *increasing convex order* $\le_{\mathrm{icx}}$, and the *majorization order* $\prec$ are defined as follows (see, e.g., Marshall and Olkin (1979), Müller and Stoyan (2002), Shaked and Shanthikumar (2007)). Given two random vectors $\boldsymbol{X}, \boldsymbol{Y}$ we say that $\boldsymbol{Y} \le_{\mathrm{st}} \boldsymbol{X}$ if

$$\mathbb{E}[\phi(\boldsymbol{Y})] \le \mathbb{E}[\phi(\boldsymbol{X})] \tag{2.1}$$

for all nondecreasing functions ϕ; we say that $\boldsymbol{Y} \le_{\mathrm{cx}} \boldsymbol{X}$ if (2.1) holds for all convex functions ϕ, and we say that $\boldsymbol{Y} \le_{\mathrm{icx}} \boldsymbol{X}$ if (2.1) holds for all nondecreasing convex functions ϕ. It is well known that $\boldsymbol{Y} \le_{\mathrm{st}} \boldsymbol{X}$ iff

$$\mathbb{P}(\boldsymbol{Y} \in A) \le \mathbb{P}(\boldsymbol{X} \in A) \quad \text{for all increasing sets } A, \tag{2.2}$$

where we call a set *increasing* if its indicator function is nondecreasing. In the case of univariate random variables X, Y, (2.2) becomes

$$\mathbb{P}(Y \le t) \ge \mathbb{P}(X \le t) \quad \text{for all } t \in \mathbb{R}.$$

The statement $\boldsymbol{X} \le_{\mathrm{st}} \boldsymbol{Y}$ depends only on the marginal laws $\mathscr{L}(\boldsymbol{X})$ and $\mathscr{L}(\boldsymbol{Y})$, so sometimes we write $\mathscr{L}(\boldsymbol{X}) \le_{\mathrm{st}} \mathscr{L}(\boldsymbol{Y})$, and analogously for $\le_{\mathrm{cx}}$ and $\le_{\mathrm{icx}}$.

Note that if X and Y are unbiased estimators of some parameter, than $X \le_{\mathrm{cx}} Y$ implies $\mathrm{Var}[X] \le \mathrm{Var}[Y]$, so that the convex order implies the more standard variance (or mean square error) comparison.

Given two vectors $\boldsymbol{x} = (x_1, \dots, x_n)$, $\boldsymbol{y} = (y_1, \dots, y_n)$, we write $\boldsymbol{y} \prec \boldsymbol{x}$ if

$$\sum_{i=1}^{k} y_{(i)} \le \sum_{i=1}^{k} x_{(i)} \quad \text{for} \quad k = 1, \dots, n-1, \qquad \sum_{i=1}^{n} y_{(i)} = \sum_{i=1}^{n} x_{(i)},$$

where $y_{(1)} \le \cdots \le y_{(n)}$ is the increasing rearrangement of $\boldsymbol{y}$, and analogously for $\boldsymbol{x}$. The relation $\boldsymbol{y} \prec \boldsymbol{x}$ holds if and only if there exists an $n \times n$ doubly stochastic matrix $\boldsymbol{D}$ such that $\boldsymbol{y} = \boldsymbol{D}\boldsymbol{x}$.

A function $\psi : \mathbb{R}^n \to \mathbb{R}$ is called Schur convex, or Schur concave, if $\boldsymbol{y} \prec \boldsymbol{x}$ implies $\psi(\boldsymbol{y}) \le \psi(\boldsymbol{x})$, or $\psi(\boldsymbol{y}) \ge \psi(\boldsymbol{x})$, respectively. If $\varphi : \mathbb{R} \to \mathbb{R}$ is convex then $\psi(\boldsymbol{x}) = \sum_{i=1}^{n} \varphi(x_i)$ is Schur convex.

A random vector $\boldsymbol{X}$ is *associated* if for all nondecreasing functions ϕ, ψ we have $\mathrm{Cov}[\phi(\boldsymbol{X}), \psi(\boldsymbol{X})] \ge 0$.

Recall that a subset $A \subset \mathbb{R}^d$ is a *lattice* if it is closed under componentwise maximum $\vee$ and minimum $\wedge$. A random vector $\boldsymbol{X}$ is *multivariate totally positive of order* 2 (MTP_2) if its support is a lattice and its density $f_{\boldsymbol{X}}$ with respect to some product measure on $\mathbb{R}^d$ satisfies

$$f_{\boldsymbol{X}}(\boldsymbol{s})\, f_{\boldsymbol{X}}(\boldsymbol{t}) \le f_{\boldsymbol{X}}(\boldsymbol{s} \vee \boldsymbol{t})\, f_{\boldsymbol{X}}(\boldsymbol{s} \wedge \boldsymbol{t}) \quad \text{for all } \boldsymbol{s}, \boldsymbol{t} \in \mathbb{R}^d.$$

MTP_2 implies association. Also, any vector having independent components is MTP_2.

Let U be a random variable with values in some measurable space $(\mathfrak{U}, \mathscr{U})$ with nonatomic law P_U. A finite sequence $\mathscr{B} = (B_1, \dots, B_b)$ of subsets of $\mathfrak{U}$ is called an *ordered partition* of $\mathfrak{U}$ if $B_i \cap B_j = \varnothing$ for $i, j \in \{1, \dots, b\}$, $i \ne j$, and $\cup_{i=1}^{b} B_i = \mathfrak{U}$. For the sake of brevity in the sequel whenever we say partition we mean ordered partition.

In this paper we consider partitions $\mathscr{B} = (B_1, \dots, B_b)$ of $\mathfrak{U}$ where the sets B_i are measurable and such that for $i = 1, \dots, b$ we have $\mathbb{P}(U \in B_i) = k_i/n$, for some $k_i \in \{1, \dots, n\}$ satisfying $\sum_i k_i = n$. We say that such a partition $\mathscr{B}$ of $\mathfrak{U}$ and a partition $\mathscr{B}^* = (B_1^*, \dots, B_b^*)$ of $N := \{1, \dots, n\}$ are associated if the cardinalities $|B_i^*|$ of the sets B_i^* satisfy $|B_i^*| = k_i$ for $i = 1, \dots, b$. We then have

$$\mathbb{P}(U \in B_i) = \frac{|B_i^*|}{n}. \tag{2.3}$$

The notation $B \in \mathscr{B}$ means that B is one of the sets B_i which comprise $\mathscr{B}$, and, given $B \in \mathscr{B}$ we let B^* denote the corresponding set B_i^* in $\mathscr{B}^*$ such that (2.3) holds.

Given two partitions $\mathscr{B}^* = (B_1^*, \dots, B_b^*)$ and $\mathscr{C}^* = (C_1^*, \dots, C_c^*)$ of N we write $\mathscr{C}^* \le_{\mathrm{ref}} \mathscr{B}^*$, that is, that $\mathscr{B}^*$ is a refinement of $\mathscr{C}^*$, when every set in $\mathscr{C}^*$ is the union of sets in $\mathscr{B}^*$. We will use the same order $\le_{\mathrm{ref}}$ also for partitions of $\mathfrak{U}$.

Clearly, if $\mathscr{C}$ and $\mathscr{B}$ are partitions of $\mathfrak{U}$, each of which can be associated to some partition of N, then when $\mathscr{C} \le_{\mathrm{ref}} \mathscr{B}$ then there exist partitions $\mathscr{C}^*$ and $\mathscr{B}^*$ associated to $\mathscr{C}$ and $\mathscr{B}$, respectively, satisfying $\mathscr{C}^* \le_{\mathrm{ref}} \mathscr{B}^*$.

Call $\mathscr{A}^* = (\{1\}, \dots, \{n\})$ the finest partition of N. Then $\mathscr{B}^* \le_{\mathrm{ref}} \mathscr{A}^*$ for all $\mathscr{B}^*$, and for any partition $\mathscr{A}$ of $\mathfrak{U}$ associated to $\mathscr{A}^*$ we have

$$\mathbb{P}(U \in A_i) = \frac{1}{n}. \tag{2.4}$$

Call $\mathscr{D}^* = (N)$ the coarsest partition of N. Then $\mathscr{D}^* \le_{\mathrm{ref}} \mathscr{B}^*$ for all $\mathscr{B}^*$.

For a partition $\mathscr{B}$ and $B \in \mathscr{B}$, let $P_{U|B}$ denote the conditional law of U given $U \in B$. Let $\{V_j^B, j \in B^*\}$ be random variables with law $P_{U|B}$ with $\{V_j^B, j \in B^*, B \in \mathscr{B}\}$ independent.

3 The supremum

Let $f : \mathfrak{U} \to \mathbb{R}$ be measurable, and define

$$W_{\mathrm{S}}^{\mathscr{B}} = \max_{B \in \mathscr{B}} \max_{j \in B^*} f(V_j^B), \tag{3.1}$$

where the subscript S indicates that $W_{\mathrm{S}}^{\mathscr{B}}$ will be used to estimate the (essential) supremum of the function f.

Given a random variable U with values in $(\mathfrak{U}, \mathscr{U})$, let $f^* := \operatorname{ess\,sup} f(U)$. It is clear that for any choice of partition $\mathscr{B}$, $\mathbb{P}(W_{\mathrm{S}}^{\mathscr{B}} \leq f^*) = 1$. The following result compares two estimators of type $W_{\mathrm{S}}^{\mathscr{B}}$. Since both estimators underestimate f^*, the stochastically larger one is preferable.

Theorem 3.1. *Let $\mathscr{C} \leq_{\mathrm{ref}} \mathscr{B}$. Then*

$$W_{\mathrm{S}}^{\mathscr{C}} \leq_{\mathrm{st}} W_{\mathrm{S}}^{\mathscr{B}}. \tag{3.2}$$

It is easy to see that $W_{\mathrm{S}}^{\mathscr{D}}$ is a consistent estimator of f^* as $n \to \infty$ when $\mathscr{D} = \{\mathfrak{U}\}$, associated to the coarsest partition $\mathscr{D}^*$. Hence Theorem 3.1 implies that $W_{\mathrm{S}}^{\mathscr{B}_n}$ is also consistent for f^* for any sequence $\mathscr{B}_n$ of partitions on $\mathfrak{U}$ which are associated to any sequence of partitions $\mathscr{B}_n^*$ of $\{1, \dots, n\}$ for $n = 1, 2, \dots$.

Though Zhigljavsky and Žilinskas (2008, Theorem 3.4) prove a similar result, we provide a short proof for completeness, depending on the following lemma.

Lemma 3.2. *Given a partition $\mathscr{B}^*$ of N, consider a collection of independent random variables $\{\xi_j^{B^*}\}$, $B^* \in \mathscr{B}^*$, $j \in B^*$, with those indexed by the same element B^* of the partition being identically distributed.*

For $\mathscr{C}^* \leq_{\mathrm{ref}} \mathscr{B}^*$ let $\{\xi_j^{C^*}\}$ with $C^* \in \mathscr{C}^*$ and $j \in C^*$ be a collection of independent random variables with the mixture distribution

$$\mathscr{L}(\xi_j^{C^*}) = \sum_{B^* \subset C^*} \frac{|B^*|}{|C^*|} \mathscr{L}(\xi_j^{B^*}). \tag{3.3}$$

Then

$$\max_{C^* \in \mathscr{C}^*} \max_{j \in C^*} \xi_j^{C^*} \leq_{\mathrm{st}} \max_{B^* \in \mathscr{B}^*} \max_{j \in B^*} \xi_j^{B^*}. \tag{3.4}$$

Proof. Let $p^{B^*} = \mathbb{P}(\xi_1^{B^*} \leq t)$ for $B^* \in \mathscr{B}^*$, and $p^{C^*} = \mathbb{P}(\xi_1^{C^*} \leq t)$ for $C^* \in \mathscr{C}^*$.

We claim that

$$(\underbrace{p^{C_1^*}, \dots, p^{C_1^*}}_{|C_1^*|}, \dots, \underbrace{p^{C_c^*}, \dots, p^{C_c^*}}_{|C_c^*|}) \prec (\underbrace{p^{B_1^*}, \dots, p^{B_1^*}}_{|B_1^*|}, \dots, \underbrace{p^{B_b^*}, \dots, p^{B_b^*}}_{|B_b^*|}).$$

To see this, observe that (3.3) implies that the vector on the left-hand side above is obtained from the one on the right by multiplying it by the $n \times n$ doubly stochastic matrix $\boldsymbol{D}$ which is block diagonal where the i-th block is the $|C_i^*| \times |C_i^*|$ matrix with all entries equal to $1/|C_i^*|$.

Hence, by the Schur concavity of the function $(\theta_1, \dots, \theta_n) \mapsto \prod_{i=1}^n \theta_i$, we have

$$\mathbb{P}\left(\max_{C^* \in \mathscr{C}^*} \max_{j \in C^*} \xi_j^{C^*} \le t\right) = \prod_{C^* \in \mathscr{C}^*} (p^{C^*})^{|C^*|} \ge \prod_{B^* \in \mathscr{B}^*} (p^{B^*})^{|B^*|} = \mathbb{P}\left(\max_{B^* \in \mathscr{B}^*} \max_{j \in B^*} \xi_j^{B^*} \le t\right),$$

which is equivalent to (3.4). □

Proof of Theorem 3.1. Let $\mathscr{B}^*$ and $\mathscr{C}^*$ be partitions associated with $\mathscr{B}$ and $\mathscr{C}$, respectively, satisfying $\mathscr{C}^* \le_{\text{ref}} \mathscr{B}^*$, and let $\{\xi_j^{B^*}, B^* \in \mathscr{B}^*, j \in B^*\}$ and $\{\xi_j^{C^*}, C^* \in \mathscr{C}^*, j \in C^*\}$ be collections of independent random variables with distributions

$$\mathbb{P}(\xi_j^{B^*} \le t) = \mathbb{P}(f(U) \le t \,|\, U \in B)$$
$$\mathbb{P}(\xi_j^{C^*} \le t) = \mathbb{P}(f(U) \le t \,|\, U \in C).$$

Then (3.3) holds (law of total probability), and the result follows by Lemma 3.2. □

Corollary 3.3. *Let $U_1, \dots, U_n$ be i.i.d. random variables with uniform distribution on $(0,1]$. Let $V_i, i = 1, \dots, n$ be independent random variables uniformly distributed on the intervals*

$$J_i = \left(\frac{i-1}{n}, \frac{i}{n}\right], \quad i = 1, \dots, n,$$

respectively. Then, for a measurable function $f : \mathfrak{U} \to \mathbb{R}$, we have

$$\max(f(U_1), \dots, f(U_n)) \le_{\text{st}} \max(f(V_1), \dots, f(V_n)).$$

Corollary 3.3 is the result that we heard from Abram Kagan. It follows from Theorem 3.1 by taking $\mathfrak{U} = (0,1]$, $P_U =$ uniform, $\mathscr{B} = (J_1, \dots, J_n)$, and $\mathscr{C} = ((0,1])$, the coarsest partition.

4 The supremum with censored observations

Let $f : \mathfrak{U} \to \mathbb{R}$ be a bounded function; without loss of generality we take $0 \le f(u) \le 1$ for all $u \in \mathfrak{U}$. In this section we assume that for a sample of points of the type $(u, t) \in \mathfrak{U} \times [0,1]$ we are allowed to observe only whether $t > f(u)$. These data may be viewed as a censored version of those in Section 3.

For any partition $\mathscr{B}$ with associated partition $\mathscr{B}^*$, let $\{V_j^B, j \in B^*\}$, $B \in \mathscr{B}$, and $\{T_j, j \in N\}$ be independent random variables with law $P_{U|B}$ and the uniform distribution on $[0,1]$, respectively, and let

$$S^{\mathscr{B}} = \bigcup_{B \in \mathscr{B}} \{j \in B^* : T_j \le f(V_j^B)\}, \quad \text{and} \quad W_{\text{CS}}^{\mathscr{B}} = \max_{j \in S^{\mathscr{B}}} T_j.$$

When $S^{\mathscr{B}} = \varnothing$ we set $W_{\text{CS}}^{\mathscr{B}} = 0$. The letter C in the subscript CS indicates censored data. Again it is clear that $\mathbb{P}(W_{\text{CS}}^{\mathscr{B}} \le f^*) = 1$, so the estimator $W_{\text{CS}}^{\mathscr{B}}$ underestimates f^*.

Theorem 4.1. *If $\mathscr{C} \leq_{\text{ref}} \mathscr{B}$, then $W^{\mathscr{C}}_{\text{CS}} \leq_{\text{st}} W^{\mathscr{B}}_{\text{CS}}$.*

Proof. Below when we write V^B_j without specifying B, we mean that $B \in \mathscr{B}$ corresponds in the sense of (2.3) to the set $B^* \in \mathscr{B}^*$ which contains the index j. For any $t \in [0,1]$ we may calculate the distribution function of $W^{\mathscr{B}}_{\text{CS}}$ at t by writing

$$\begin{aligned}\{W^{\mathscr{B}}_{\text{CS}} \leq t\} &= \bigcup_{R\subset N} \left\{\max_{j\in S^{\mathscr{B}}} T_j \leq t, S^{\mathscr{B}} = R\right\} \\ &= \bigcup_{R\subset N} \left\{T_j \leq t, T_j \leq f(V^B_j) \text{ for all } j \in R, \text{ and } T_j > f(V^B_j) \text{ for all } j \notin R\right\} \\ &= \bigcup_{R\subset N} \left\{T_j \leq t \wedge f(V^B_j) \text{ for all } j \in R, \text{ and } T_j > f(V^B_j) \text{ for all } j \notin R\right\}.\end{aligned}$$

Hence, conditionally on $\{V^B_j,\, j \in B^*,\, B \in \mathscr{B}\}$, we obtain, using the fact that T_j are uniform:

$$\begin{aligned}\mathbb{P}(W^{\mathscr{B}}_{\text{CS}} \leq t \,|\, V^B_j, j \in B^*, B \in \mathscr{B}) &= \sum_{R\subset N} \prod_{j\in R} \mathbb{P}(T_j \leq t \wedge f(V^B_j)) \prod_{j\notin R} \mathbb{P}(T_j > f(V^B_j)) \\ &= \sum_{R\subset N} \prod_{j\in R} (t \wedge f(V^B_j)) \prod_{j\notin R} (1 - f(V^B_j)) \qquad (4.1) \\ &= \sum_{h_1=1}^{|B^*_1|} \cdots \sum_{h_b=1}^{|B^*_b|} \sum_{\substack{R\subset N \\ |R\cap B^*_i|=h_i}} \prod_{j\in R} (t \wedge f(V^B_j)) \prod_{j\notin R} (1 - f(V^B_j)).\end{aligned}$$

Taking expectation we obtain the unconditional distribution,

$$\begin{aligned}\mathbb{P}(W^{\mathscr{B}}_{\text{CS}} \leq t) = &\sum_{h_1=1}^{|B^*_1|} \cdots \sum_{h_b=1}^{|B^*_b|} \prod_{i=1}^{b} \binom{|B^*_i|}{h_i} \left(\int_{B_i} (t \wedge f(u))\, \mathrm{d}P_{U|B_i}(u)\right)^{h_i} \\ &\cdot \left(\int_{B_i} (1 - f(u))\, \mathrm{d}P_{U|B_i}(u)\right)^{|B^*_i|-h_i} \\ = &\prod_{B\in\mathscr{B}} \left(\int_B (t \wedge f(u))\, \mathrm{d}P_{U|B}(u) + \int_B (1 - f(u))\, \mathrm{d}P_{U|B}(u)\right)^{|B^*|}.\end{aligned}$$

Let

$$q^B = \int_B (t \wedge f(v))\, \mathrm{d}P_{U|B}(v) + \int_B (1 - f(v))\, \mathrm{d}P_{U|B}(v) = \int_B [(t \wedge f(v)) + (1 - f(v))]\, \mathrm{d}P_{U|B}(v).$$

If C is a union of disjoint sets B_i then

$$q^C = \sum_i q^{B_i} \frac{\mathbb{P}(B_i)}{\mathbb{P}(C)}.$$

As in the proof of Lemma 3.2, if $\mathscr{C} \leq_{\text{ref}} \mathscr{B}$ then

$$(\underbrace{q^{C_1}, \ldots, q^{C_1}}_{|C_1^*|}, \ldots, \underbrace{q^{C_c}, \ldots, q^{C_c}}_{|C_c^*|}) \prec (\underbrace{q^{B_1}, \ldots, q^{B_1}}_{|B_1^*|}, \ldots, \underbrace{q^{B_b}, \ldots, q^{B_b}}_{|B_b^*|}).$$

Therefore, again by Schur concavity of the function $(\theta_1, \ldots, \theta_n) \mapsto \prod_{i=1}^n \theta_i$, we have

$$\mathbb{P}(W_{\text{CS}}^{\mathscr{B}} \leq t) = \prod_{B \in \mathscr{B}} \left(q^B\right)^{|B^*|} \geq \prod_{C \in \mathscr{C}} \left(q^C\right)^{|C^*|} = \mathbb{P}(W_{\text{CS}}^{\mathscr{C}} \leq t).$$

$\square$

It is easy to see that $W_{\text{CS}}^{\mathscr{D}}$ is consistent as $n \to \infty$. Therefore, as for $W_{\text{S}}^{\mathscr{B}}$, Theorem 4.1 implies that $W_{CS}^{\mathscr{B}_n}$ is consistent as $n \to \infty$ for any sequence $\mathscr{B}_n$ associated with partitions $\mathscr{B}_n^*$ of $\{1, \ldots, n\}$.

5 The integral

With the subscript I standing for integral, let

$$W_{\text{I}}^{\mathscr{B}} = \frac{1}{n} \sum_{B \in \mathscr{B}} \sum_{j \in B^*} f(V_j^B) \tag{5.1}$$

be the estimator for $\overline{f} := \mathbb{E}[f(U)] = \int f(U) \, \mathrm{d}\mathbb{P}$ when $\int |f(U)| \, \mathrm{d}\mathbb{P}$ is finite.

Note that

$$\begin{aligned} \mathbb{E}[W_{\text{I}}^{\mathscr{B}}] &= \frac{1}{n} \sum_{B \in \mathscr{B}} \sum_{j \in B^*} \mathbb{E}[f(U) \,|\, U \in B] = \sum_{B \in \mathscr{B}} \frac{|B^*|}{n} \mathbb{E}[f(U) \,|\, U \in B] \\ &= \sum_{i=1}^{b} \mathbb{P}(U \in B_i) \, \mathbb{E}[f(U) \,|\, U \in B_i] = \mathbb{E}[f(U)], \end{aligned}$$

demonstrating that $W_{\text{I}}^{\mathscr{B}}$ is an unbiased estimator of $\overline{f}$. Our first result compares two such unbiased estimators in terms of variance, or, equivalently, mean square error.

Theorem 5.1. *Let $\mathscr{C} \leq_{\text{ref}} \mathscr{B}$. Then*

$$\text{Var}[W_{\text{I}}^{\mathscr{B}}] \leq \text{Var}[W_{\text{I}}^{\mathscr{C}}].$$

Theorem 5.1 is a particular case of Theorem 5.2, which considers situations where the function f at any point $u \in \mathfrak{U}$ can only be observed with random mean zero error. In such cases, let

$$W_{\text{IE}}^{\mathscr{B}} = \frac{1}{n} \sum_{B \in \mathscr{B}} \sum_{j \in B^*} \left(f(V_j^B) + \varepsilon_j\right), \tag{5.2}$$

where the variables ε_j are independent copies of a random variable ε having mean 0 and finite variance, and are independent of the variables V_j^B. Clearly, as $W_{\mathrm{IE}}^{\mathscr{B}}$ differs from $W_{\mathrm{I}}^{\mathscr{B}}$ by a mean zero error, $W_{\mathrm{IE}}^{\mathscr{B}}$ is unbiased for $\overline{f}$, and $W_{\mathrm{I}}^{\mathscr{B}}$ is the special case of $W_{\mathrm{IE}}^{\mathscr{B}}$ when the error has zero variance. In particular, Theorem 5.2 below implies Theorem 5.1.

Theorem 5.2. *Let $\mathscr{C} \leq_{\mathrm{ref}} \mathscr{B}$. Then*

$$\mathrm{Var}[W_{\mathrm{IE}}^{\mathscr{B}}] \leq \mathrm{Var}[W_{\mathrm{IE}}^{\mathscr{C}}].$$

Proof. In what follows we consider conditional expectation with respect to a partition. Though the notion is standard, specifically, by $\mathbb{E}[f(U) + \varepsilon \,|\, \mathscr{B}]$ we mean the random variable that takes values $\overline{f}_B := \mathbb{E}[f(U) \,|\, U \in B]$ with probability $|B^*|/n$. Then

$$\begin{aligned}\mathrm{Var}[f(U) + \varepsilon \,|\, \mathscr{B}] &= \mathbb{E}\left[\{f(U) + \varepsilon - \mathbb{E}[f(U) + \varepsilon \,|\, \mathscr{B}]\,\}^2 \,|\, \mathscr{B}\right] \\ &= \mathbb{E}\left[\{f(U) + \varepsilon - \mathbb{E}[f(U) \,|\, \mathscr{B}]\}^2 \,|\, \mathscr{B}\right]\end{aligned}$$

is a random variable taking values $\mathbb{E}\left[\left(f(U) + \varepsilon - \overline{f}_B\right)^2 \,|\, U \in B\right]$ with probability $|B^*|/n$, and

$$\begin{aligned}\mathbb{E}[\mathrm{Var}[f(U) + \varepsilon \,|\, \mathscr{B}]] &= \sum_{B \in \mathscr{B}} \frac{|B^*|}{n} \mathbb{E}\left[(f(U) + \varepsilon - \overline{f}_B)^2 \,|\, U \in B\right] \\ &= \frac{1}{n} \sum_{B \in \mathscr{B}} |B^*|\, \mathbb{E}\left[(f(V_1^B) + \varepsilon - \overline{f}_B)^2\right] \\ &= \frac{1}{n} \mathrm{Var}\left[\sum_{B \in \mathscr{B}} \sum_{j \in B_i^*} f(V_j^B) + \varepsilon_j^B\right] \\ &= n\, \mathrm{Var}[W_{\mathrm{IE}}^{\mathscr{B}}].\end{aligned}$$

If $\mathscr{C} \leq_{\mathrm{ref}} \mathscr{B}$, then for any random variable Y, say, $\mathrm{Var}[\mathbb{E}[Y \,|\, \mathscr{B}]] \geq \mathrm{Var}[\mathbb{E}[Y \,|\, \mathscr{C}]]$ by Jensen's inequality, and now the usual variance decomposition of Y (aka the Pythagorean Theorem) implies $\mathbb{E}[\mathrm{Var}[Y \,|\, \mathscr{B}]] \leq \mathbb{E}[\mathrm{Var}[Y \,|\, \mathscr{C}]]$. Therefore

$$\mathbb{E}[\mathrm{Var}[f(U) + \varepsilon \,|\, \mathscr{B}]] \leq \mathbb{E}[\mathrm{Var}[f(U) + \varepsilon \,|\, \mathscr{C}]],$$

and hence

$$\mathrm{Var}[W_{\mathrm{IE}}^{\mathscr{B}}] = \frac{1}{n}\mathbb{E}[\mathrm{Var}[f(U) + \varepsilon \,|\, \mathscr{B}]] \leq \frac{1}{n}\mathbb{E}[\mathrm{Var}[f(U) + \varepsilon \,|\, \mathscr{C}]] = \mathrm{Var}[W_{\mathrm{IE}}^{\mathscr{C}}].$$

$\square$

It follows immediately from Theorem 5.1 that $\mathrm{Var}[W_{\mathrm{I}}^{\mathscr{A}}] \leq \mathrm{Var}[W_{\mathrm{I}}^{\mathscr{D}}]$. The following counterexample shows nevertheless that, even when the function is observed without error, $W_{\mathrm{I}}^{\mathscr{A}} \not\leq_{\mathrm{cx}} W_{\mathrm{I}}^{\mathscr{D}}$, that is, that domination in the convex order does not hold. In the counterexample we consider the absolute (L_1) rather than mean square error (L_2).

Example 5.3. Let $\mathfrak{U} = [0,1]$ and U have a uniform distribution on $[0,1]$. Furthermore let $n = 2$, and $A_1 = [0, 1/2], A_2 = (1/2, 1]$. Define

$$f(u) = 4I_{[0,1/2]}(u) + 2I_{(1/2,3/4]}(u) + 6I_{(3/4,1]}(u).$$

Then $W_{\mathrm{I}}^{\mathscr{D}}$ takes the values $2,3,4,5,6$ with probabilities $(1,4,6,4,1)/16$, respectively. The variable $W_{\mathrm{I}}^{\mathscr{A}}$, based on one random observation from each of the above intervals A_i, takes the values 3 and 5 each with probability $1/2$. Therefore $\mathbb{E}[W_{\mathrm{I}}^{\mathscr{A}}] = 4 = \mathbb{E}[W_{\mathrm{I}}^{\mathscr{D}}]$.

We have $\mathrm{Var}[W_{\mathrm{I}}^{\mathscr{D}}] = \mathrm{Var}[W_{\mathrm{I}}^{\mathscr{A}}] = 1$, but for the convex function $\psi(u) = |u-4|$ we have

$$\mathbb{E}[\psi(W_{\mathrm{I}}^{\mathscr{D}})] = \mathbb{E}|W_{\mathrm{I}}^{\mathscr{D}} - 4| = 2\frac{2}{16} + 2\frac{4}{16} = \frac{12}{16} < 1 = \mathbb{E}|W_{\mathrm{I}}^{\mathscr{A}} - 4| = \mathbb{E}[\psi(W_{\mathrm{I}}^{\mathscr{A}})].$$

6 The integral with censored observation

Keeping the notation and the spirit of Section 4, assume that for a sample of points of the type $(u,t) \in \mathfrak{U} \times [0,1]$ we are allowed to observe only whether $t \le f(u)$, and let

$$W_{\mathrm{CI}}^{\mathscr{B}} = \frac{1}{n} \sum_{B \in \mathscr{B}} \sum_{j \in B^*} I_{\{T_j \le f(V_j^B)\}}.$$

Note that $W_{\mathrm{CI}}^{\mathscr{B}}$ is an unbiased estimator of $\overline{f} = \mathbb{E}[f(U)]$, as

$$\begin{aligned}
\mathbb{E}[W_{\mathrm{CI}}^{\mathscr{B}}] &= \frac{1}{n} \sum_{B \in \mathscr{B}} \sum_{j \in B^*} \mathbb{P}(T_j \le f(V_j^B)) = \frac{1}{n} \sum_{B \in \mathscr{B}} \sum_{j \in B^*} \int_{\mathfrak{U}} \int_0^1 I_{\{t \le f(u)\}} \,\mathrm{d}t \,\mathrm{d}P_{U|B}(u) \\
&= \sum_{B \in \mathscr{B}} \frac{|B^*|}{n} \int_{\mathfrak{U}} f(u) \,\mathrm{d}P_{U|B}(u) = \sum_{B \in \mathscr{B}} \mathbb{P}(B) \mathbb{E}[f(U) | U \in B] \\
&= \mathbb{E}[f(U)].
\end{aligned}$$

Theorem 6.1. *Let $\mathscr{C} \le_{\mathrm{ref}} \mathscr{B}$. Then*

$$W_{\mathrm{CI}}^{\mathscr{B}} \le_{\mathrm{cx}} W_{\mathrm{CI}}^{\mathscr{C}}.$$

Proof. By a result in Karlin and Novikoff (1963) (see also Marshall and Olkin (1979, Theorems 12 and 15.E)), if

$$X_{\boldsymbol{p}} = \frac{1}{n} \sum_{i=1}^n \xi_i,$$

where $\xi_1, \ldots, \xi_n$ are independent Bernoulli variables with parameters $p_1, \ldots, p_n$, and $\boldsymbol{p} = (p_1, \ldots, p_n)$, then

$$\boldsymbol{p} \prec \boldsymbol{q} \quad \text{implies} \quad X_{\boldsymbol{q}} \le_{\mathrm{cx}} X_{\boldsymbol{p}}. \tag{6.1}$$

Define

$$p^B = \mathbb{P}(T_j \le f(V_j^B)), \quad p^C = \mathbb{P}(T_j \le f(V_j^C)),$$

and

$$\boldsymbol{p} = (\underbrace{p^{C_1}, \dots, p^{C_1}}_{|C_1^*|}, \dots, \underbrace{p^{C_c}, \dots, p^{C_c}}_{|C_c^*|}), \quad \boldsymbol{q} = (\underbrace{p^{B_1}, \dots, p^{B_1}}_{|B_1^*|}, \dots, \underbrace{p^{B_b}, \dots, p^{B_b}}_{|B_b^*|}).$$

If $C = \bigcup_i B_i$ then

$$p^C = \sum_i p^{B_i} \frac{|B_i|}{|C|},$$

so $\boldsymbol{p} \prec \boldsymbol{q}$ and invoking (6.1) completes the proof. □

Notice that in the case of censored observations the comparison is stronger than in the case of perfect observations, since it holds in the convex order and not just for the variance.

7 Convex order under monotonicity

Though Example 5.3 shows that the convex order does not hold in general between estimators $W_{\mathrm{I}}^{\mathscr{B}}$ and $W_{\mathrm{I}}^{\mathscr{C}}$ when $\mathscr{C} \le_{\mathrm{ref}} \mathscr{B}$, Theorem 7.1 below shows that the convex order does hold for certain partitions when $\mathfrak{U}$ is a totally ordered space and the function f is monotone. In Subsection 7.2 we extend these results to multivariate functions.

7.1 Univariate functions

In the rest of this subsection the space $\mathfrak{U}$ is totally ordered, and without loss of generality we choose $\mathfrak{U} = [0, 1]$. For subsets G and H of the real line, we write $G \le H$ if $g \le h$ for every $g \in G$ and $h \in H$. We call a partition $\mathscr{B} = (B_1, \dots, B_b)$ of $\mathfrak{U}$ monotone if $B_1 \le \cdots \le B_b$.

Theorem 7.1. *Let $\mathscr{B}$ and $\mathscr{C}$ be monotone partitions of $\mathfrak{U}$ and let $\mathscr{C} \le_{\mathrm{ref}} \mathscr{B}$. If f is nondecreasing, then*

$$W_{\mathrm{I}}^{\mathscr{B}} \le_{\mathrm{cx}} W_{\mathrm{I}}^{\mathscr{C}}. \tag{7.1}$$

Theorem 7.1 is a special case of the following theorem.

Theorem 7.2. *Let $\mathscr{B}$ and $\mathscr{C}$ be monotone partitions of $\mathfrak{U}$ and let $\mathscr{C} \le_{\mathrm{ref}} \mathscr{B}$. If f is nondecreasing, then*

$$W_{\mathrm{IE}}^{\mathscr{B}} \le_{\mathrm{cx}} W_{\mathrm{IE}}^{\mathscr{C}}. \tag{7.2}$$

To prove Theorem 7.2 we will apply the following lemma.

Lemma 7.3. *Let ξ and η be random variables such that $\xi \leq_{\mathrm{st}} \eta$, and let ξ_i and η_j be independent copies of ξ and η respectively. Let K be an integer valued random variable, independent of all ξ_j and η_j, satisfying $K \leq m$ for some integer m, and having an integer valued expectation, $\mathbb{E}[K] = k$. Then*

$$\sum_{j=1}^{k} \xi_j + \sum_{j=k+1}^{m} \eta_j \leq_{\mathrm{cx}} \sum_{j=1}^{K} \xi_j + \sum_{j=K+1}^{m} \eta_j. \tag{7.3}$$

Proof. Since $\xi \leq_{\mathrm{st}} \eta$ we may construct i.i.d. pairs (ξ_i, η_i) with $\mathbb{P}(\xi_i \leq \eta_i) = 1$ for all $i = 1, \dots, m$. We adopt the usual convention that if $k = 0$ then $\sum_{j=1}^{k} \xi_j = 0$. First note that, by Wald's Lemma,

$$\mathbb{E}\left[\sum_{j=1}^{k} \xi_j + \sum_{j=k+1}^{m} \eta_j\right] = \mathbb{E}\left[\sum_{j=1}^{K} \xi_j + \sum_{j=K+1}^{m} \eta_j\right].$$

Therefore (see, e.g., Müller and Stoyan (2002, Theorem 1.5.3)) it suffices to show that

$$\sum_{j=1}^{k} \xi_j + \sum_{j=k+1}^{m} \eta_j \leq_{\mathrm{icx}} \sum_{j=1}^{K} \xi_j + \sum_{j=K+1}^{m} \eta_j.$$

Let ϕ be an increasing convex function and set

$$g(k) := \mathbb{E}\left[\phi\left(\sum_{j=1}^{k} \xi_j + \sum_{j=k+1}^{m} \eta_j\right)\right].$$

Note that

$$g(k) = \mathbb{E}\left[\phi\left(\sum_{j=1}^{K} \xi_j + \sum_{j=K+1}^{m} \eta_j\right) \Big| K = k\right] \quad \text{and} \quad \mathbb{E}[g(K)] = \mathbb{E}\left[\phi\left(\sum_{j=1}^{K} \xi_j + \sum_{j=K+1}^{m} \eta_j\right)\right].$$

Thus we have to show that $g(k) \leq \mathbb{E}[g(K)]$. Since $\mathbb{E}[K] = k$, this follows readily by Jensen's inequality, once we prove that $g(k)$ is a convex function.

The following part of the proof follows ideas of Ross and Schechner (1984). Setting

$$S_k = \sum_{j=1}^{k} \xi_j + \sum_{j=k+2}^{m} \eta_j,$$

We have

$$g(k+1) - g(k) = \mathbb{E}[\phi(\xi_{k+1} + S_k)] - \mathbb{E}[\phi(\eta_{k+1} + S_k)].$$

Since ϕ is convex, and $\xi_{k+1} \leq \eta_{k+1}$, the function

$$h(s) := \mathbb{E}[\phi(\xi_{k+1} + S_k) \,|\, S_k = s] - \mathbb{E}[\phi(\eta_{k+1} + S_k) \,|\, S_k = s]$$

is decreasing in s. Now note that

$$S_{k+1} = \sum_{i=1}^{k+1} \xi_i + \sum_{i=k+3}^{m} \eta_i \leq_{\mathrm{st}} S_k = \sum_{i=1}^{k} \xi_i + \sum_{i=k+2}^{m} \eta_i$$

because $\xi_{k+1} \leq_{\mathrm{st}} \eta_{k+2}$. Hence $g(k+1) - g(k) = \mathbb{E}[h(S_k)]$ is increasing in k, thus proving that g is convex, as required. □

Proof of Theorem 7.2. Since $\mathscr{B} = (B_1, \dots, B_b)$ and $\mathscr{C} = (C_1, \dots, C_c)$ are monotone partitions satisfying $\mathscr{C} \leq_{\mathrm{ref}} \mathscr{B}$ there exist $1 = i_1 < i_2 < \cdots < i_c < i_{c+1} = b+1$ such that

$$C_q = \bigcup_{j=i_q}^{i_{q+1}-1} B_j, \quad \text{for} \quad q = 1, \dots, c.$$

As the union above may be formed by taking the union of two consecutive sets at a time, it suffices to prove (7.2) for the case where $c = b-1$, $C_m = B_m \cup B_{m+1}$, $C_k = B_k$ for $k \in \{1, \dots, m-1\}$, and $C_k = B_{k+1}$ for $k \in \{m+1, \dots, c\}$.

In this case we have

$$W_{\mathrm{IE}}^{\mathscr{B}} = \frac{1}{n} \left[\sum_{C \neq C_m} \sum_{j \in C^*} f(V_j^C) + \sum_{j \in B_m^*} f(V_j^{B_m}) + \sum_{j \in B_{m+1}^*} f(V_j^{B_{m+1}}) + \sum_{j \in N} \varepsilon_j \right],$$

$$W_{\mathrm{IE}}^{\mathscr{C}} = \frac{1}{n} \left[\sum_{C \neq C_m} \sum_{j \in C^*} f(V_j^C) + \sum_{j \in C_m^*} f(V_j^{C_m}) + \sum_{j \in N} \varepsilon_j \right].$$

Note that

$$\mathscr{L}\left(\sum_{j \in C_m^*} f(V_j^{C_m}) \right) = \mathscr{L}\left(\sum_{j=1}^{K} f(V_j^{B_m}) + \sum_{j=K+1}^{|C_m^*|} f(V_j^{B_{m+1}}) \right),$$

where K is distributed according to a binomial with parameters

$$\left(|C_m^*|, \frac{|B_m^*|}{|C_m^*|} \right).$$

It is easy to see that if two variables are ordered by the convex order (see (2.1)) and we add the same independent variable to each one, to wit, $\sum_{j \in N} \varepsilon_j$, then the convex order is preserved. This fact and Lemma 7.3 now yield (7.2). □

Note that the the variance of the left hand side variable in (7.3) is easily seen to be smaller than that of the variable on the right by the usual variance decomposition, without assuming $\xi \leq_{\mathrm{st}} \eta$. This leads to another (but not very different) proof of Theorem 5.1 using the arguments in the proof of Theorem 7.2.

We end this section with a natural extension of Lemma 7.3 which we think is of independent interest. As it can be proved from Lemma 7.3 by straightforward induction, the details are omitted. We need the following notation: $\mathscr{K}_{mn}$ denotes the class of vectors $\boldsymbol{k} = (k_1, \dots, k_n)$ with nonnegative integer components such that $\sum_{i=1}^n k_i = m$.

Proposition 7.4. *Let ξ_i be random variables satisfying $\xi_1 \le_{\mathrm{st}} \xi_2 \le_{\mathrm{st}} \dots \le_{\mathrm{st}} \xi_n$, and let $\{\xi_{ij}\}_{i=1,\dots,n,\, j=1,2,\dots}$ be independent random variables with $\mathscr{L}(\xi_{ij}) = \mathscr{L}(\xi_i)$. Given $\boldsymbol{k} \in \mathscr{K}_{mn}$, let $\boldsymbol{K} = (K_1, \dots, K_n)$ be a random vector having a multinomial distribution with parameters $(m, \boldsymbol{k}/m)$, independent of $\{\xi_{ij}\}$. Then*

$$\sum_{i=1}^{n} \sum_{j=1}^{k_i} \xi_{ij} \le_{\mathrm{cx}} \sum_{i=1}^{n} \sum_{j=1}^{K_i} \xi_{ij}. \tag{7.4}$$

7.2 Multivariate functions

In this section we extend the results in Section 7.1 for monotone functions observed with noise on a sample of points, to the multivariate case. When we consider multivariate monotone functions, stratifying can still yield improvement but some restrictions are necessary, both on the distribution of the random vector used for sampling and on the stratifying partitions. More specifically, we consider estimation of an integral or expectation with respect to a random vector whose components exhibit independence or positive dependence as defined below, and stratification that preserves such dependence in the sense that it continues to hold for the random vector when conditioned on each set of the partition.

Let $f : [0,1]^d \to [0,1]$ be nondecreasing in each variable, and let $\boldsymbol{U}$ be a random vector taking values in $[0,1]^d$ with a nonatomic distribution. Our goal is to show that the estimate of $\mathbb{E}[f(\boldsymbol{U})]$ improves by refining stratifications as follows: Start with a partition $\mathscr{C} = (C_1, \dots, C_b)$ of $[0,1]^d$ such that for each i the distribution $\mathscr{L}(\boldsymbol{U} \,|\, \boldsymbol{U} \in C_i)$ is associated. Then split C_i into $C_i \cap G$ and $C_i \cap G^c$, where G is an increasing set. Lemma 7.7 below shows that the new partition obtained by this splitting achieves a better estimator of the integral in terms of the convex order, and Theorem 7.5 provides some necessary conditions for its application.

Theorem 7.5. *Consider a partition $\mathscr{C} = (C_1, \dots, C_c)$ of $[0,1]^d$ where each C_i is a lattice. Let $\mathscr{B}$ be a partition obtained by a sequence of refinements $\mathscr{C} = \mathscr{C}_1 \le_{\mathrm{ref}} \dots \le_{\mathrm{ref}} \mathscr{C}_m = \mathscr{B}$, such that for $k = 1, \dots, m-1$ the partition $\mathscr{C}_{k+1}$ is obtained from $\mathscr{C}_k$ by splitting one set of $\mathscr{C}_k$, say $C_{i_k,k}$, into $C_{i_k,k} \cap G_k$ and $C_{i_k,k} \cap G_k^c$, where G_k, G_k^c are lattices, and G_k is an increasing set.*

If $\boldsymbol{U}$ is MTP$_2$ on $[0,1]^d$ and $f : [0,1]^d \to [0,1]$ is nondecreasing, then

$$W_{\mathrm{IE}}^{\mathscr{B}} \le_{\mathrm{cx}} W_{\mathrm{IE}}^{\mathscr{C}}.$$

The proof of Theorem 7.5 is preceded by the following lemmas.

Lemma 7.6. *If $\boldsymbol{U}$ is an associated random vector, and G is an increasing set, then*

$$\mathscr{L}(\boldsymbol{U} \mid \boldsymbol{U} \in G^c) \leq_{\mathrm{st}} \mathscr{L}(\boldsymbol{U} \mid \boldsymbol{U} \in G). \tag{7.5}$$

Conversely, if (7.5) *holds for every increasing set G, then $\boldsymbol{U}$ is associated.*

Proof. First note that (7.5) is equivalent to

$$\mathbb{P}(\boldsymbol{U} \in A \mid \boldsymbol{U} \in G) \geq \mathbb{P}(\boldsymbol{U} \in A \mid \boldsymbol{U} \in G^c)$$

holding for all increasing sets A. The latter inequality is easily seen to be equivalent to

$$\mathbb{P}(\boldsymbol{U} \in A \cap G)[1 - \mathbb{P}(\boldsymbol{U} \in G)] \geq [\mathbb{P}(\boldsymbol{U} \in A) - \mathbb{P}(\boldsymbol{U} \in A \cap G)]\mathbb{P}(\boldsymbol{U} \in G).$$

By simple cancelation this is equivalent to

$$\mathbb{P}(\boldsymbol{U} \in A \cap G) \geq \mathbb{P}(\boldsymbol{U} \in A)\mathbb{P}(\boldsymbol{U} \in G),$$

which is equivalent to association of the random vector $\boldsymbol{U}$ by e.g., Shaked (1982). □

Lemma 7.7. *Consider a partition $\mathscr{C} = (C_1, \ldots, C_c)$ of $[0,1]^d$ such that for some C_i the distribution $\mathscr{L}(\boldsymbol{U} \mid \boldsymbol{U} \in C_i)$ is associated. Let G be an increasing set and let $\mathscr{B} = (C_1, \ldots, C_{i-1}, C_i \cap G, C_i \cap G^c, C_{i+1}, \ldots, C_c)$. If $f : [0,1]^d \to [0,1]$ is nondecreasing, then*

$$W_{\mathrm{IE}}^{\mathscr{B}} \leq_{\mathrm{cx}} W_{\mathrm{IE}}^{\mathscr{C}}.$$

Proof. With $\mathscr{L}(\boldsymbol{V}_1) = \mathscr{L}(\boldsymbol{U} \mid \boldsymbol{U} \in C_i \cap G^c)$ and $\mathscr{L}(\boldsymbol{V}_2) = \mathscr{L}(\boldsymbol{U} \mid \boldsymbol{U} \in C_i \cap G)$, Lemma 7.6 yields $\boldsymbol{V}_1 \leq_{\mathrm{st}} \boldsymbol{V}_2$. Letting $\xi = f(\boldsymbol{V}_1) + \varepsilon_1$ and $\eta = f(\boldsymbol{V}_2) + \varepsilon_2$, with $\varepsilon_1, \varepsilon_2$ i.i.d., the monotonicity of f implies $\xi \leq_{\mathrm{st}} \eta$, and Lemma 7.3 now proves the claim, applying arguments as in the proof of Theorem 7.2. □

The following result can be found in Karlin and Rinott (1980).

Lemma 7.8. *If an MTP_2 vector $\boldsymbol{U}$ takes values in a lattice, of which C is a sublattice, then $\mathscr{L}(\boldsymbol{U} \mid \boldsymbol{U} \in C)$ is MTP_2 and hence associated.*

The following corollary is obvious, and only requires the fact that the intersection of sublattices is a lattice.

Corollary 7.9. *If an MTP_2 vector $\boldsymbol{U}$ takes values in some lattice, and C, G and G^c, are all sublattices, then both $\mathscr{L}(\boldsymbol{U} \mid \boldsymbol{U} \in C \cap G)$ and $\mathscr{L}(\boldsymbol{U} \mid \boldsymbol{U} \in C \cap G^c)$ are MTP_2, and hence also associated.*

Proof of Theorem 7.5. We first prove by induction that $\mathscr{L}(\boldsymbol{U} \mid \boldsymbol{U} \in C_{i,k})$ are MTP_2 for all $C_{i,k} \in \mathscr{C}_k$ and $k = 1, \dots, m$. For $k = 1$ this follows by Lemma 7.8 and the assumptions that $\boldsymbol{U}$ is MTP_2 and that $C_i = C_{i,1}$ are sublattices of $[0,1]^d$. Assuming the statement true for $1 \le k < m$, to verify that it is true for $k+1$ we need only show that $\mathscr{L}(\boldsymbol{U} \mid \boldsymbol{U} \in C_{i_k,k} \cap G_k)$ and $\mathscr{L}(\boldsymbol{U} \mid \boldsymbol{U} \in C_{i_k,k} \cap G_k^c)$ are MTP_2, which follows from Lemma 7.8, thus completing the induction.

Hence, again using by Lemma 7.8, $\mathscr{L}(\boldsymbol{U} \mid \boldsymbol{U} \in C_{i_k,k})$ is associated. Since G_k is increasing, Lemma 7.7 now yields

$$W_{\mathrm{IE}}^{\mathscr{C}_{k+1}} \le_{\mathrm{cx}} W_{\mathrm{IE}}^{\mathscr{C}_k} \quad \text{for all} \quad k = 1, \dots, m-1,$$

and, therefore, the theorem. $\square$

A natural example where Theorem 7.5 may be applied is the case where the lattices C_i are boxes of the form $\{\boldsymbol{x} = (x_1, \dots, x_d) \in [0,1]^d \,:\, a_i \le x_i \le b_i,\ i = 1, \dots, d\}$, and the increasing sets are specified as $G = \{\boldsymbol{x} = (x_1, \dots, x_d) \in [0,1]^d \,:\, a \le x_j\}$ for some j. In particular one can start with the partition which consists of the whole space $[0,1]^d$ and split it into boxes by repeatedly choosing a partition element to subdivide by taking its intersection with some such G and G^c. In $[0,1]^2$ the resulting partition forms a tiling of the square by rectangles. Note that from the first step, a sequence of partitions created using G as above has at least one line which crosses the whole square from side to side. Therefore the tiling of Figure 1 is not attainable by such a sequence.

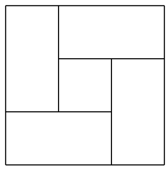

Figure 1

Lastly, we notice that the hypothesis of MTP_2 includes as a particular case the uniform distribution on $[0,1]^d$, so Theorem 7.5 applies to the estimation of the integral $\int f(\boldsymbol{u})\, \mathrm{d}\boldsymbol{u}$ on $[0,1]^d$, or any lattice.

8 Symmetric sampling designs

The main purpose of this section is to show that the convex order between unbiased estimators of $\mathbb{E}[f(U)]$, for a monotone f, can hold even when the partitions determining the estimators are not ordered by refinement. We start with the following definition and lemmas.

Definition 8.1. Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be random vectors in $\mathbb{R}^n$. We say that $\boldsymbol{X} \prec^{\mathrm{sc}} \boldsymbol{Y}$ if $\mathbb{E}[g(\boldsymbol{X})] \le \mathbb{E}[g(\boldsymbol{Y})]$ for every Schur convex function g.

Since an increasing function of a Schur convex function is also Schur convex, it is easy to see that $\boldsymbol{X} \prec^{\mathrm{sc}} \boldsymbol{Y}$ is equivalent to $g(\boldsymbol{X}) \leq_{\mathrm{st}} g(\boldsymbol{Y})$ for every Schur convex function g (see Nevius et al. (1977)).

The following result is a very special case of a well-known result of Strassen (1965) (see also Alfsen (1971) and Lindvall (2002)).

Lemma 8.2. *Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be n-dimensional random vectors with compact support such that $\boldsymbol{X} \prec^{\mathrm{sc}} \boldsymbol{Y}$. Then there exists a coupling of $(\boldsymbol{X}, \boldsymbol{Y})$ such that $\mathbb{P}(\boldsymbol{X} \prec \boldsymbol{Y}) = 1$, that is, a coupling where the vector $\boldsymbol{X}$ is majorized by $\boldsymbol{Y}$ with probability 1.*

Lemma 8.3. *Let $\xi_1, \dots, \xi_n$ be random variables satisfying $\xi_1 \leq \xi_2 \leq \dots \leq \xi_n$ a.s., and let $\{\xi_{ij}\}_{i=1,\dots,n,\, j=1,2,\dots}$ be independent with $\mathscr{L}(\xi_{ij}) = \mathscr{L}(\xi_i)$. Consider two vectors having nonnegative integer components $\boldsymbol{\ell} = (\ell_1, \dots, \ell_n)$ and $\boldsymbol{k} = (k_1, \dots, k_n)$ such that $\boldsymbol{\ell} \prec \boldsymbol{k}$. Let Π be a uniformly distributed permutation in the permutation group of $\{1, \dots, n\}$, independent of $\{\xi_{ij}\}$. Define*

$$Z_{\boldsymbol{\ell}} = \sum_{i=1}^{n} \sum_{j=1}^{\ell_{\Pi(i)}} \xi_{ij} \quad \textit{and} \quad Z_{\boldsymbol{k}} = \sum_{i=1}^{n} \sum_{j=1}^{k_{\Pi(i)}} \xi_{ij}. \tag{8.1}$$

Then

$$Z_{\boldsymbol{\ell}} \leq_{\mathrm{cx}} Z_{\boldsymbol{k}}$$

Proof. It is well a known fact in majorization (see, e.g., Hardy et al. (1952, Proof of Lemma 2, p. 47) and Marshall and Olkin (1979, Lemma B.1, p. 21)) that to prove the lemma it suffices to consider $\boldsymbol{k}$ and $\boldsymbol{\ell}$ that differ in only two coordinates, and moreover, it is easy to see that it suffices to consider the special case where $\boldsymbol{k}$ and $\boldsymbol{\ell}$ satisfy

$$k_1 - 1 \geq k_2 + 1, \quad \ell_1 = k_1 - 1, \quad \ell_2 = k_2 + 1, \quad \ell_i = k_i, \text{ for } i = 3, \dots, n. \tag{8.2}$$

Write the desired conclusion as

$$\sum_{i:\Pi(i)\in\{1,2\}} \sum_{j=1}^{\ell_{\Pi(i)}} \xi_{ij} + \sum_{i:\Pi(i)\notin\{1,2\}} \sum_{j=1}^{\ell_{\Pi(i)}} \xi_{ij} \leq_{\mathrm{cx}} \sum_{i:\Pi(i)\in\{1,2\}} \sum_{j=1}^{k_{\Pi(i)}} \xi_{ij} + \sum_{i:\Pi(i)\notin\{1,2\}} \sum_{j=1}^{k_{\Pi(i)}} \xi_{ij}.$$

Since $\ell_{\Pi(i)} = k_{\Pi(i)}$ for all $\Pi(i) \notin \{1,2\}$, it suffices to prove

$$\sum_{j=1}^{\ell_1} \xi_{\Pi^{-1}(1),j} + \sum_{j=1}^{\ell_2} \xi_{\Pi^{-1}(2),j} \leq_{\mathrm{cx}} \sum_{j=1}^{k_1} \xi_{\Pi^{-1}(1),j} + \sum_{j=1}^{k_2} \xi_{\Pi^{-1}(2),j}.$$

The distributions of the sums above are mixtures over random permutations. By pairing a permutation Π in this mixture with the permutation Γ for which $\Gamma^{-1}(1) =$

$\Pi^{-1}(2), \Gamma^{-1}(2) = \Pi^{-1}(1)$ and $\Gamma^{-1}(i) = \Pi^{-1}(i)$ for all $i \notin \{1,2\}$, and relabeling $\Pi^{-1}(i)$ as i for $i \in \{1,2\}$, we see that it suffices to show

$$\frac{1}{2}\mathscr{L}\left(\sum_{j=1}^{\ell_1}\xi_{1j} + \sum_{j=1}^{\ell_2}\xi_{2j}\right) + \frac{1}{2}\mathscr{L}\left(\sum_{j=1}^{\ell_2}\xi_{1j} + \sum_{j=1}^{\ell_1}\xi_{2j}\right) \leq_{\mathrm{cx}} \frac{1}{2}\mathscr{L}\left(\sum_{j=1}^{k_1}\xi_{1j} + \sum_{j=1}^{k_2}\xi_{2j}\right) + \frac{1}{2}\mathscr{L}\left(\sum_{j=1}^{k_2}\xi_{1j} + \sum_{j=1}^{k_1}\xi_{2j}\right).$$

By (8.2), the above is equivalent to

$$\frac{1}{2}\mathscr{L}\left(\sum_{j=1}^{k_1-1}\xi_{1j} + \sum_{j=1}^{k_2}\xi_{2j} + \xi_{2k_1}\right) + \frac{1}{2}\mathscr{L}\left(\sum_{j=1}^{k_2}\xi_{1j} \ + \ \xi_{1k_1} \ + \sum_{j=1}^{k_1-1}\xi_{2j}\right) \leq_{\mathrm{cx}} \frac{1}{2}\mathscr{L}\left(\sum_{j=1}^{k_1}\xi_{1j} + \sum_{j=1}^{k_2}\xi_{2j}\right) + \frac{1}{2}\mathscr{L}\left(\sum_{j=1}^{k_2}\xi_{1j} + \sum_{j=1}^{k_1}\xi_{2j}\right). \quad (8.3)$$

We claim that $(a_1, a_2) \prec (b_1, b_2)$ for

$$\begin{aligned}(a_1, a_2) &= \left(\sum_{j=1}^{k_1-1}\xi_{1j} + \sum_{j=1}^{k_2}\xi_{2j} + \xi_{2k_1}, \sum_{j=1}^{k_2}\xi_{1j} + \xi_{1k_1} + \sum_{j=1}^{k_1-1}\xi_{2j}\right), \\ (b_1, b_2) &= \left(\sum_{j=1}^{k_1}\xi_{1j} + \sum_{j=1}^{k_2}\xi_{2j}, \sum_{j=1}^{k_2}\xi_{1j} + \sum_{j=1}^{k_1}\xi_{2j}\right).\end{aligned}$$

To prove $(a_1, a_2) \prec (b_1, b_2)$ first note that $a_1 + a_2 = b_1 + b_2$, so it suffices to show that $\max\{a_1, a_2\} \leq b_2$, which is a consequence of

$$b_2 - a_1 = \sum_{j=k_2+1}^{k_1-1}(\xi_{2j} - \xi_{1j}) \geq 0, \quad \text{and} \quad b_2 - a_2 = \xi_{2k_1} - \xi_{1k_1} \geq 0.$$

Recalling that if φ convex on $\mathbb{R}$ then $\sum \varphi(x_i)$ is Schur convex, it follows that for any convex function φ we have $\varphi(a_1) + \varphi(a_2) \leq \varphi(b_1) + \varphi(b_2)$, and (8.3) follows readily. □

The next proposition, where we replace the random permutation in Lemma 8.3 by an assumption of exchangeability, is of possible interest by itself. Let

$$\Upsilon_{\boldsymbol{\ell}} = \sum_{i=1}^{n}\sum_{j=1}^{\ell_i}\xi_{ij}$$

Proposition 8.4. *Let $\xi_1 \le \xi_2 \le \ldots \le \xi_n$ a.s., and let $\{\xi_{ij}\}_{i=1,\ldots,n,\, j=1,2,\ldots}$ be independent with $\mathscr{L}(\xi_{ij}) = \mathscr{L}(\xi_i)$. Let $\boldsymbol{L}$ and $\boldsymbol{K}$ be nonnegative and bounded integer valued exchangeable random vectors, independent of $\{\xi_{ij}\}$. If $\boldsymbol{L} \prec^{\mathrm{sc}} \boldsymbol{K}$, then*

$$\Upsilon_{\boldsymbol{L}} \le_{\mathrm{cx}} \Upsilon_{\boldsymbol{K}}. \tag{8.4}$$

Proof. Since $\boldsymbol{K}$ is exchangeable $\mathscr{L}(Z_{\boldsymbol{K}}) = \mathscr{L}(\Upsilon_{\boldsymbol{K}})$ where $Z_{\boldsymbol{k}}$ is defined in (8.1), so it suffices to show $Z_{\boldsymbol{L}} \le_{\mathrm{cx}} Z_{\boldsymbol{K}}$. By Lemma 8.2 we may take $\boldsymbol{L} \prec \boldsymbol{K}$ almost surely, and now, using the assumed independence, Lemma 8.3 may be invoked to complete the argument. □

Recall that $\mathscr{K} = \mathscr{K}_{nn}$ is the class of vectors $\boldsymbol{k} = (k_1, \ldots, k_n)$ with nonnegative integer components such that $\sum_{i=1}^n k_i = n$, and let $\mathbf{1}_n := (1, \ldots, 1) \in \mathbb{R}^n$.

Corollary 8.5. *Let $\xi_1 \le \xi_2 \le \ldots \le \xi_n$ a.s., and let $\{\xi_{ij}\}_{i=1,\ldots,n,\, j=1,2,\ldots}$ be independent with $\mathscr{L}(\xi_{ij}) = \mathscr{L}(\xi_i)$. If $\boldsymbol{K}$ is an exchangeable random vector in $\mathscr{K}_{nn}$, independent of $\{\xi_{ij}\}$, then*

$$\Upsilon_{\mathbf{1}_n} \le_{\mathrm{cx}} \Upsilon_{\boldsymbol{K}}.$$

These results will allow us to compare sampling designs for the estimation of $\mathbb{E}[f(U)]$ on $\mathfrak{U}$ which are not refinements of one another. As in Theorem 5.2, we assume that for any v in a sample of points on $\mathfrak{U}$ we observe $f(v) + \varepsilon$ where ε is a mean zero independent random error.

Definition 8.6. Given an exchangeable random vector $\boldsymbol{K} = (K_1, \ldots, K_n) \in \mathscr{K}$ and a partition $\mathscr{A} = (A_1, \ldots, A_n)$ of $\mathfrak{U}$ such that $\mathbb{P}(U \in A_i) = 1/n$ for $i = 1, \ldots, n$, the associated *symmetric random design* $\mathbb{K}$ is the design consisting of independent subsamples $V_{i1}, \ldots, V_{iK_i}$ of size K_i with distribution $P_{U|A_i}$, $i = 1, \ldots, n$.

By symmetry and the fact that the error ε has mean zero, it is easy to see that the estimator

$$W_{\mathbb{K}} = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{K_i} f(V_{ij}) + \varepsilon_{ij} \tag{8.5}$$

is unbiased for $\mathbb{E}[f(U)]$.

Given a nonrandom $\boldsymbol{k} \in \mathscr{K}$ let $\boldsymbol{K}$ be the exchangeable random vector having the uniform distribution over all permutations of $\boldsymbol{k}$, and let $\mathbb{K}$ be the associated symmetric random design. In other words, a sample having the design $\mathbb{K}$ can be realized by choosing a random permutation Π of $(1, \ldots, n)$, and then sampling $k_{\Pi(i)}$ observations from $P_{U|A_i}$ for all i. Note that $\mathbb{1}_n$, the associated symmetric random design of $\mathbf{1}_n$, is identical to the design discussed in earlier sections of sampling one observation from each subset of a partition $\mathscr{A} = (A_1, \ldots, A_n)$ of $\mathfrak{U}$ such that $\mathbb{P}(U \in A_i) = 1/n$ for $i = 1, \ldots, n$.

Theorem 8.7. *Let $\mathfrak{U}$ be a totally ordered set, f be a nondecreasing function and $\mathscr{A} = (A_1, \ldots, A_n)$ a monotone partition of $\mathfrak{U}$ satisfying $\mathbb{P}(U \in A_i) = 1/n$. Consider exchangeable vectors $\boldsymbol{L}, \boldsymbol{K} \in \mathscr{K}$ satisfying $\boldsymbol{L} \prec^{\mathrm{sc}} \boldsymbol{K}$, and let $\mathbb{L}$ and $\mathbb{K}$ be their associated symmetric random designs. Then*

$$W_{\mathbb{L}} \leq_{\mathrm{cx}} W_{\mathbb{K}}. \tag{8.6}$$

In particular (8.6) *holds for the random designs $\mathbb{L}$ and $\mathbb{K}$ associated respectively to some fixed $\boldsymbol{\ell}$ and $\boldsymbol{k}$ satisfying $\boldsymbol{\ell} \prec \boldsymbol{k}$. Moreover, for any symmetric random design $\mathbb{K}$ we have*

$$W_{\mathbb{1}_n} \leq_{\mathrm{cx}} W_{\mathbb{K}}.$$

Proof. Recalling that V_{ij} is the j-th sampled value in partition element A_i, the theorem follows directly by applying Proposition 8.4 to $\xi_{ij} = f(V_{ij}) + \varepsilon_{ij}$. $\square$

Note that Theorem 8.7 allows comparisons of designs which are not ordered by refinement. For example, for $n = 4$, we can compare the symmetric designs $\mathbb{L}$ and $\mathbb{K}$ associated with $\boldsymbol{\ell} = (2, 2, 0, 0)$, and $\boldsymbol{k} = (3, 1, 0, 0)$. For the design $\mathbb{L}$ one chooses two sets at random from the finest partition, and takes samples of size of 2 from each. For $\mathbb{K}$ one again begins by choosing two sets at random from the finest partition, but then takes samples of sizes 3 and 1 from the sets chosen. As the sets in which to sample chosen by these two designs may differ, the associated partitions may not be ordered by refinement. Yet, as $\boldsymbol{\ell} \prec \boldsymbol{k}$, Theorem 8.7 allows us to conclude that $W_{\mathbb{L}}$ is smaller in the convex order than $W_{\mathbb{K}}$.

The notion of symmetric design involves additional randomization and is different from the kind of sampling treated before. Such symmetric randomization is required in order to preserve unbiasedness, a property shared by all the estimators of integrals considered here. With randomization, it is seen that more balanced designs, in the sense defined by the majorization order, are better.

The last part of Theorem 8.7 shows that for a sample of size n, a partition $\mathscr{A}$ into equal probability subsets with a sample of one from each is best in the sense of convex order. However, such precise sampling may not always be feasible. Lastly, we note this result is consistent with the inequality given in Theorem 7.2, which also confirms that higher levels of stratification are preferable.

References

ALFSEN, E. M. (1971) *Compact Convex Sets and Boundary Integrals.* Springer-Verlag, New York.

HARDY, G. H., LITTLEWOOD, J. E., and PÓLYA, G. (1952) *Inequalities.* Cambridge, at the University Press. 2d ed.

Karlin, S. and Novikoff, A. (1963) Generalized convex inequalities. *Pacific J. Math.* **13**, 1251–1279.

Karlin, S. and Rinott, Y. (1980) Classes of orderings of measures and related correlation inequalities. I. Multivariate totally positive distributions. *J. Multivariate Anal.* **10**, 467–498.

Lindvall, T. (2002) *Lectures on the Coupling Method.* Dover Publications Inc., Mineola, NY. Corrected reprint of the 1992 original.

Marshall, A. W. and Olkin, I. (1979) *Inequalities: Theory of Majorization and Its Applications.* Academic Press Inc., New York.

Müller, A. and Stoyan, D. (2002) *Comparison Methods for Stochastic Models and Risks.* John Wiley & Sons Ltd., Chichester.

Nevius, S. E., Proschan, F., and Sethuraman, J. (1977) Schur functions in statistics. II. Stochastic majorization. *Ann. Statist.* **5**, 263–273.

Novak, E. (1988) *Deterministic and Stochastic Error Bounds in Numerical Analysis*, volume 1349 of *Lecture Notes in Mathematics.* Springer-Verlag, Berlin.

Ross, S. M. and Schechner, Z. (1984) Some reliability applications of the variability ordering. *Oper. Res.* **32**, 679–687.

Shaked, M. (1982) A general theory of some positive dependence notions. *J. Multivariate Anal.* **12**, 199–218.

Shaked, M. and Shanthikumar, J. G. (2007) *Stochastic Orders.* Springer Series in Statistics. Springer, New York.

Strassen, V. (1965) The existence of probability measures with given marginals. *Ann. Math. Statist* **36**, 423–439.

Zhigljavsky, A. and Žilinskas, A. (2008) *Stochastic Global Optimization.* Springer, New York.

Zhigljavsky, A. A. and Chekmasov, M. V. (1996) Comparison of independent, stratified and random covering sample schemes in optimization problems. *Math. Comput. Modelling* **23**, 97–110.

April 22, 2009